PP: Mining a 'Trove': Modeling a Transnational Literary Culture

This paper describes a project that uses automatic search and harvesting of a digitized historical newspaper database to create an archive of serial fiction in those newspapers. My focus is the Trove database of Australian newspapers, created by the National Library of Australia. But this approach to the archive is not only applicable to Australian newspapers or to fiction. To emphasize this wider applicability I'll touch only briefly on specific results. Instead, I'll focus in the first part of the paper on our methodology, which has important applications for research in bibliography, literary studies and book history. The second part of the paper will focus on two key epistemological principles the project raises: one, the necessity of understanding the cultural objects we explore and analyze as multidimensional; and two, the necessity of conceptualizing the archive as provisional.

It's long been recognized that much of the fiction read by nineteenth-century Australians was published in newspapers. But we've had very little idea what fiction was available due to the size of the archive. With hundreds of newspapers – many containing multiple serialisations per edition – a systematic manual search for fiction has not been feasible. The possibilities for searching this archive have dramatically expanded with the creation of the National Library of Australia's Trove database.

PP: Trove is the largest collection of digitized historical newspapers internationally, with its 13.5 million pages from over 680 newspapers far surpassing the 7.8 million pages in Chronicling America and the 8.2 million pages in the British Newspaper Archive (those figures are from late June). And unlike the British Newspaper Archive, it's freely available.

This project doesn't search Trove for particular titles and authors, an approach that would tend to discover what's already assumed to be present. Rather, we're leveraging the generic forms of newspapers by searching for words or phrases commonly used to frame serial fiction in the nineteenth-century Australian press. I should say, when I say "we" in this paper I'm referring to myself and Carol Hetherington, who's a bibliographer employed full time on this project for three years with funding from the Australian Research Council.

The first search term we've trialled is "chapter", and it's proven effective in optimising results for fiction for two reasons. First, the word often occurs multiple times in the text designated by Trove as an "article" (because a single instalment often contains many chapters); and second, because the word frequently appears in the "article" title (which is defined as the first four lines, and is the only part of the text manually checked and transcribed, thus reducing the effects of OCR errors on search results). As the project continues we'll use other search terms, including "serial story", "our storyteller" and "new novelist". Each will have its own benefits and drawbacks, and our aim is to employ a range of terms until the returned results demonstrate a high level of repetition with what is already indexed. We'll then sample newspapers to gauge how successful this method has been in identifying fiction in the digitised record. We'll also make the full archive available and editable online.

The search term "chapter" returned more than 800,000 results. We exported metadata and full-text files for the first 250,000 using an API (or Application Programming

Interface) created by Tim Sherratt, now the manager of Trove.

PP: The metadata exported is on the left, and includes a unique identifier for the article, the first four lines of text (the "title"), dates and pages for publication, information about the newspaper, permanent urls for the newspaper page and the "article", and a count of crowd-sourced corrections on the OCR text.

Due to the usefulness of "chapter" in optimising the relevance ranking for fiction, we found that the first 30 or so sets of 5000 results were almost exclusively fiction, with the share of other records to fiction increasing over the next 20 sets of 5000 results. We only looked at the first 50 sets of 5000 because, after that point, the proportion of fiction found was not significant enough to warrant us continuing (in other words, for our purposes the relevance ranking algorithm had exhausted its usefulness). Other results of the "chapter" search not relevant to our project include reports of meetings of a chapter of a lodge or religious association, accounts of a chapter in the life of a town or person, or even public documents such as deeds of grant and regulations organised in chapter divisions.

After removing duplicates (of which there were many) and any records we did not consider fiction we've added a range of bibliographical fields (listed here on the right), some relating to the presentation of fiction in the newspapers, and some based on further bibliographical research, and I'll discuss, toward the end of the paper, why we consider so many bibliographical fields to be necessary. Despite considerable automation, and the fact that it's often possible to generalise across multiple records with substantially similar "titles", this is a time-consuming process: Carol has been working on the "chapter" search for about a year. However, we believe this first search will be the most time-consuming (because future searches will include many of the same results). And although time-consuming, this process has enabled us to uncover a huge amount of fiction in Australian newspapers.

PP For the nineteenth century this process has yielded:

- 58,717 unique records (or instalments – and remember, we also have the full text for these)
- these instalments constitute 6,269 titles
- of which 4,076 are unique (as you can see in the difference between the number of titles and the number of unique titles, many stories are published multiple times in different newspapers – and even, in some cases, in the same newspaper, a decade or so apart).

Many of the authors published anonymously, or used pseudonyms or signatures only. We've been able to identify:

- 1,693 individual authors of these titles;
- there remain 1,301 titles by authors we have not yet been able to identify

As I said, this is only the first of a number of searches. Its results show the potential of this method to enrich our bibliographic record and, in the process, to demonstrate and leverage the potential of major digitization projects.

We can use this dataset to explore multiple research questions, but as I said, I want to focus on general principles, so I'll only give one quick example relating to the nationality and gender of authors. It's been assumed that the fiction published in

nineteenth-century Australian newspapers essentially replicated the fiction serialized in Britain. In fact, it turns out that more Australian than British titles were serialized from the mid-1840s to the mid-1860s and, more surprisingly (given the technological developments that had occurred by that time), from the mid-1870s to the end of the 1880s. Accordingly, while it's been assumed that the fiction serialized in Australia – as in Britain and American – was predominantly by women, in fact, the majority is by men. I'm referring, here, not only to local fiction, but to overseas titles as well. This finding suggests that, when sourcing fiction from overseas, Australian newspaper editors specifically sought out less common men's writing for their Australian readers. That is, unless we're mistaken about the gender dynamics in American and British serial fiction: because, after all, our understanding of the contents of those newspaper archives is largely based on individual examples, contemporaneous anecdote and sampling of particular (usually "small") magazines.

So, as I said, a very quick example; and there's many more questions that could be asked, using both bibliometric and text-mining approaches. This work will build on existing data-led studies, most notably, Franco Moretti's "distant reading" and Matt Jockers' "macroanalysis". But it will have two major strengths not present in these other projects. First, all the data we collect – including the full texts – will be made publicly available so that other scholars can explore, check, extend, refine, and potentially challenge, any findings; and so that this data can be reused in future research. I'd like to emphasise this point because I think it's absolutely imperative for digital humanities research to prioritise the availability of data. One questioner in a paper yesterday raised the importance of a digital humanities "data dump", and I would certainly support that initiative. Second, where these other projects explore general corpora, typically relating to the nationality of authors, this project explores works published and read at a particular time and place, and thus foregrounds specific historical and cultural context.

In the time I've got remaining I'd like to discuss two broad epistemological principles that this project foregrounds. The first is the necessity of perceiving the cultural objects we represent and study in digital humanities as multidimensional. While I'll focus on literary works, or print cultural objects, the principle applies broadly.

As some of you will know too well, nineteenth-century newspaper publishing was chaotic and unregulated. In addition to anonymous and pseudonymous publication, works were often reprinted multiple times with different titles or attributions. We've found cases with up to eight different titles accorded to substantially the same text. And as fiction moved across national borders in the nineteenth century, it was often plagiarised, rewritten and localised. For instance, American author "Old Sleuth's" novel *The American Detective in Russia* is serialised in several Australian newspapers as "Barnes, the Australian Detective".

This mutability is particularly apparent in nineteenth century newspaper publishing. But it is not peculiar to it. As Jerome McGann argues in his recent book, *A New Republic of Letters*, literary works – like all cultural objects – are constituted by their histories of production and reception. Because these histories are constantly evolving, so too is the nature and meaning, the definition and constitution, of those objects. At the most simple level, many literary works are published multiple times. Although we know all of these publication events by the name of the "work", what constitutes the

work changes over time, rendering it a process rather than a stable object.

For much of the twentieth century, literary studies has ignored the processual nature of literary works in its focus on *the* "text". Unfortunately, much digital and data-led literary research perpetuates this framework, treating literary works as if they are singular and stable. We can see this approach in bibliometric studies that represent the literary work via a single date or place of publication; and in text mining studies that analyse literary works using a single "text".

To move away from a one-dimensional conception of our objects of study, when creating an archive or collection based on mining digitised documents, the bibliographic complexity and multidimensionality of print culture should be at the forefront of our mind. It is for this reason that our project foregrounds the collection of bibliographic data: as a recognition that the meaning of literary works does not simply reside in a single, textual instantiation.

My sense is that a focus on "thick data" as well as "big data" might help us with what Alan Liu has called the "meaning problem" in digital humanities. By this, he means the failure of the field to move from data to interpretation and argument. I wonder if one of the reasons for this problem – at least in digital literary studies – is that the datasets we're "reading" aren't complicated – or thick – enough. In approaching literary works as singular stable data points or "texts" (and in the process jettisoning all original paratext while ignoring the one we create in remediating that textual object) we miss multiple dimensions of their existence and meanings.

Concentrating on the multidimensionality of literary works and the thickness of the cultural archive might help us move beyond this meaning problem. In particular, I'm finding that, because this dataset embeds multiple instantiations of the work within a complex set of bibliographical relationships, it's possible to read the collection almost like one reads a literary works. That is to say, rather than looking at a visualisation and attempting to interpret it, I am able to follow archival threads in ways that highlight particular aspects of the literary works relationship to its print cultural context.

The second epistemological principle that this project foregrounds is the provisionality of an archive produced through digital methods. Of course, all archival research is provisional, in that we never access the full documentary record, whether because it's been misplaced, mislabelled, damaged, lost or destroyed. But digital methods and resources increase the potential for unrealised disjunctions between our perceived – and our achieved – access to the archive. They do this, I think, by increasing the proxies or models involved in the research process. Archival research always involves proxies or models. In respect to literary studies, at least for the period since the invention of the printing press, each document in the archive represents – or is a proxy for – many other documents (such as the many newspapers of the same title and date, the vast majority of which no longer exist). Archival research is also mediated by a particular collection model, such as a library card catalogue.

Digital methods and resources introduce many additional proxies into these processes of representation and access. In our project, the csv and full text data we extract are models for the OCR rendered digitised newspaper pages collected by Trove, which in

turn, are models of the newspapers—or of the microfiche models of those newspapers—held in library archives throughout Australia. Trove's search interface not only provides a system for accessing these documentary proxies, but mediates such access via multiple other models. For instance, each search result or collection "view" "has its own home page, its own relevance ranking algorithm, and its own facets, influenced by the type of material included in the view". While the emergence and growing significance of digital archives does not introduce proxies and models into archival research, it does amplify the significance of these mediating factors, and render as a matter of urgency an explicit theorisation of their role and effects.

As the number of proxies or models multiplies, so too does the potential for any changes in these representational systems, or any biases or errors they introduce or perpetuate, to alter how the collection is accessed and represented; and to do so in a complex, cascading way. Some changes or problems – such as the digitisation of additional documents, or biases in the search results because certain titles aren't digitised, or even the effects of OCR errors – are relatively easy to identify and accommodate. The same cannot be said of other processes that construct, and frame our access to, the archive. In particular, the effects of search algorithms and interfaces can be difficult to perceive. Sampling may reveal specific ways in which these technologies misrepresent or omit aspects of the archive. But one can never be certain that all issues have been dealt with because, without these devices, our only access to the archive is manual, returning us to difficulties presented by its sheer size.

The importance of emphasising provisionality in respect to digital projects is increased by the rhetoric of objectivity and comprehensiveness that attends such research, where the very scale of information recovered, as well as the supposedly direct access to an underlying system provided by computers, encourage a perception that all has been revealed. In terms of theorising this process, we've been talking for a long time in digital humanities about modelling as an iterative process. What this project suggests to me is that this iterative process – controlled and executed by the researcher – is only one layer of modelling on top of a vast number of other models, some of which cannot be seen, that constantly inform and alter each other.

From this perspective, at the same time as they massively increase and enhance our access to the archive, digital resources and methods introduce additional uncertainties into this process of knowing. In that sense, I'm saying something a bit different from Bruno Latour, the other night, in that he's arguing that digital humanities exposes the traces of what we have always done or had or been, but have been unable to see. In this exposure, it enables us to better understand the processes by which we come to know, and thus, to know better and to communicate better with other disciplines. I think this is true to a significant extent, but digital methods also amplify existing features of our research practices, or even add new layers, requiring us to also investigate, understand and represent these traces: if we can.

Ultimately, rather than trying to conceal or deny uncertainties in the research process – or even to pretend that we can fully measure and constrain it – digital humanities should accept a conception of data as provisional and perpetual. What we know is constantly changing by virtue of its status as cultural objects, and we can never know all there is to know about these. Rather than replacing rigor or method, uncertainty emphasises their importance, because these are the only ways we can progress: by

gradually thickening what we know and engaging us in a continual process of reflecting on how it is we know it.